第34卷 第10期 计 算 机 工 程 2008年5月
Vol.34 No.10 Computer Engineering May 2008
·人工智能及识别技术· 文章编号：1000—3428(2008)10—0161—02 文献标识码：A 中图分类号：TP301.6

基于动态排位信息的语音关键词确认方法

陈玉平，韩纪庆，郑铁然
（哈尔滨工业大学计算机科学与技术学院，哈尔滨 150001）

摘 要：给出一种适用于在线垃圾模型的基于动态排位信息的关键词确认方法，利用识别过程中声学得分的排位信息进行关键词确认，能在不降低检出率的同时有效降低系统的误警率，效果优于同类方法。该方法不依赖于具体的关键词表，计算简单，能够应用于实际工程中。
关键词：语音识别；关键词检出；在线垃圾模型；关键词确认

Speech Keyword Verification Based on Dynamic Ranking Information

CHEN Yu-ping, HAN Ji-qing, ZHENG Tie-ran
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001)

【Abstract】 This paper proposes a novel keyword verification algorithm based on dynamic ranking information for on-line garbage model keyword spotting systems, using ranking information of acoustics scores in the recognition process. It reduces false alarm rate effectively without lowering detection rate and gets better verification performance than others. The method does not depend on specific keywords vocabulary, is easy to calculate and can be applied to practical projects.
【Key words】 speech recognition; Keyword Spotting(KWS); on-line garbage model; keyword verification

1 概述

关键词检出(Keyword Spotting, KWS)是连续语音识别的一个重要分支，有着广阔的应用前景，如国防监听、电话接听、智能家居系统、口语识别等，已成为近年来颇受重视的一个研究方向。用隐马尔可夫模型(Hidden Markov Models, HMM)对无限制语音进行关键词检出已获得了重大进展。通常，关键词检出系统引入垃圾(garbage)模型来增强关键词和非关键词的区分能力。垃圾模型的建立有 2 种方法：离线(off-line)的垃圾模型和在线(on-line)的垃圾模型。离线垃圾模型能够比较精细地刻画词表外词的特性，但必须经过精心的设计和训练，难度较大。并且离线建模使得垃圾模型的设计和训练依赖于词表的内容，当关键词发生变化时，模型需要重新训练。在线垃圾建模方法没有明确地建立垃圾模型，不依赖具体的关键词表，识别速度快，具有灵活性和稳定性的优点，更易于应用到实际工程中。但是，在线垃圾建模方法虚警率比较高，有效的关键词确认方法对于提高在线垃圾模型关键词检出系统性能非常关键。

关键词确认是关键词检出系统的重要步骤。通常，实用化的关键词检出系统分成 2 个阶段：识别阶段和确认阶段。系统在识别阶段为了保证最终结果有较高检出率，常常给出尽可能多的候选，以便包含正确的候选，因此，确认单元必须使用有效的方法拒识错误的候选，以降低系统的虚警率，同时也要保证检出率不受影响。

现有的许多关键词确认方法利用经过单独训练的确认模型来估计正反 2 种假设的分布，如文献[1]使用反关键词模型确认数字串语音。这些方法需要针对拒识进行额外的建模和训练，计算复杂，而且与关键词表密切相关，当关键词表变化时，需要重新训练，不适合在线垃圾模型关键词检出系统的确认。使用识别过程中的信息进行关键词确认可以解决这个问题，如基于长度归一化的声学置信度的方法[2]、使用音素模型的N-best平均分数作为反关键词得分的方法[3]。

基于使用识别过程中的信息来进行关键词确认的思想，本文将每帧语音在识别过程中对所有模型的声学得分进行由大到小排序，并利用它在关键词模型下的声学得分排位的归一化值来判断其是否是关键词，提出了基于动态排位信息的关键词确认方法，确认性能优于文献[2-3]的方法。

2 在线垃圾模型

本文的关键词检出系统是基于在线垃圾模型的。在线垃圾建模方法[3]不是明确地建立垃圾模型，而是在识别单位是音素时，在线地计算音素模型局部每一帧的得分，将N个最好得分的平均作为垃圾似然得分。对在线垃圾模型，关键词模型一般使用上下文无关(Context-independent, CI)或上下文相关(Context-dependent, CD)音素模型来描述。这种计算垃圾评分方法与用垃圾语料训练离线垃圾模型一样是一种平滑技术。离线垃圾模型方法使用训练样本的全局信息，而在线垃圾评分方法使用在线的、局部的信息。此外这种方法有一定的抗噪性，在噪声环境下，关键词得分发生变化，垃圾得分也跟随同方向地变化，在一定程度上起到凸显关键词语音的作用[4]。

3 基于动态排位信息的关键词确认

假设一段 T 帧的语音 $O=\{o_1,o_2,\cdots,o_T\}$，在识别过程中被认为是关键词 Kw，其模型为 Λ_{Kw}。识别网络由 K 个模型 $M=\{\Lambda_k;k=1,2,\cdots,K\}$ 组成，$\Lambda_{Kw}\in M$。在用 Viterbi 算法对这

基金项目：国家“973”计划基金资助项目(2007CB311104)
作者简介：陈玉平(1982－)，男，硕士，主研方向：语音关键词检出；韩纪庆，教授、博士生导师；郑铁然，讲师
收稿日期：2007-06-20 **E-mail：**yup.chen@utstar.com

段语音进行解码时，解码的结果是语音的每一帧对应于模型中的一个状态，每个模型内部的状态排列就称为此模型的状态序列。设 o_t 对应于 $N(o_t)$ 个模型 $M(o_t)=\{\Lambda_j;1\le j\le N(o_t)\}$，$\Lambda_{Kw}\in M(o_t)$ 的一个状态，且对于 Λ_j，o_t 对应于状态 $S(\Lambda_j,o_t)$，相应的似然得分值为

$$L_j(o_t)=\ln P(o_t\mid S(\Lambda_j,o_t)),\ \ j=1,2,\cdots,N(o_t)) \tag{1}$$

将所有 $L_j(o_t)$ 从大到小排列为

$$L_{j_1}(o_t)>L_{j_2}(o_t)>\cdots>L_{j_k}(o_t)>\cdots>L_{j_{N(o_t)}}(o_t) \tag{2}$$

其中，关键词模型 Λ_{Kw} 对应似然得分值 $L_{Kw}(o_t)=L_{j_k}(o_t)$，在式(2)中排在第 k 位，则在 o_t 帧上的动态排位为 k，在帧级(frame level)上的动态排位信息得分为 $k/N(o_t)$，写成表达式的形式为

$$Q(o_t\mid\Lambda_{Kw})=\frac{\sum_{k=1}^{N(o_t)}G(L_k(o_t)-L_{Kw}(o_t))}{N(o_t)} \tag{3}$$

其中，

$$G(L_k(o_t)-L_{Kw}(o_t))=\begin{cases}0 & \text{if } L_k(o_t)\le L_{Kw}(o_t)\\ 1 & \text{others}\end{cases} \tag{4}$$

将帧级的动态排位信息得分整合到词级上，即在整个语音段的动态排位信息得分，就是每一帧动态排位信息得分的累加和，且要对语音帧长做归一化处理。

词级(word level)的动态排位得分为

$$CM(O\mid\Lambda_{Kw})=\frac{1}{T}\sum_{t=1}^{T}Q(o_t\mid\Lambda_{Kw}) \tag{5}$$

根据动态排位信息得分 $CM(O\mid\Lambda_{Kw})$ 来进行关键词确认的准则为：(1)若 $CM(O\mid\Lambda_{Kw})<threshold$，则判断为关键词 Kw；(2)若 $CM(O\mid\Lambda_{Kw})>threshold$，则判断为虚警，舍弃。其中，$threshold$ 为拒识门限，可依据实验结果统计得到。对于所有关键词可以设置统一的拒识门限。

式(1)中的似然得分值已在Viterbi 解码过程中计算得到，动态排位信息得分增加的计算量只有式(3)中$Q(o_t)$的计算以及式(5)中的归一化，只增加$O(KT)$个加减法和T+1 个除法，此外需要一个长度为K的数组来储存临时数据$L_j(o_t)$(j=1, 2,⋯, $N(o_t)$)，系统增加的负担很小。

动态排位信息只依赖于识别过程中的似然得分值，关键词表的变化对动态排位信息得分几乎没有影响，基于动态排位信息的关键词确认方法具有较好的稳定性，特别适合关键词表经常变化的实际应用。若语音段存在噪声，将同时影响关键词模型和解码过程中其他模型的似然得分值，对它们的排位影响相对较小，因此，基于动态排位信息的关键词确认方法还具有一定的抗噪性。

4 实验结果及分析

本文训练语料采用了国家“863”语料库(共 90 821 句)，测试语料选用了微软公司发布的测试语料，共 500 句话，每句话 3 s~6 s，总时长 44 min16 s。全部数据用 16 kHz 采样，取自自然发音、实验室环境。

4.1 基线系统

建立基于在线垃圾模型的关键词检出系统，声学模型采用 5 状态连续隐马尔可夫模型(Continuous Hidden Markov Models, CHMM)，识别参数是 39 维的特征矢量，分别是 12 维的Mel频率倒谱系数(Mel Frequency Cepstrum Coefficient, MFCC)、1 维能量及它们的一阶、二阶差分。选择经济、信息等 10 个常出现的词作为关键词，在测试语料中共出现 111 次。关键词模型采用上下文相关的音素模型，与所有的音素模型并行组成识别网络，使用Viterbi beam搜索算法解码，在解码后形成的Lattice中搜索候选关键词[5]。

系统具有较高的实时性，检出率高达 97.3%，若采用 n-best 搜索算法，检出率还可以进一步提高。但虚警率比较高，占候选关键词的 71.5%。因此，寻找一种有效的关键词确认方法十分重要。实验对基于动态排位信息的关键词确认方法与基于长度归一化声学置信度的关键词确认方法、基于在线垃圾得分的关键词确认方法进行了比较。

4.2 关键词检出系统性能评价

系统的性能评价是个非常重要的问题，在统一的评价标准之下可以对研究方案进行比较，从而不断提高研究的水平。主要的性能评价标准有检出率、误警率、ROC 曲线、FOM (Figure of Merit)等。ROC 曲线为关键词的检测正确率 Pd 与误警率之间的关系曲线 FOM 为某一特定的误警率范围(通常为每小时语料每个关键词 0~10 个误警数)内关键词的平均检测正确率，这一性能评价标准被大多数关键词识别研究者所认同和接受。本文采用 ROC 曲线和 FOM 评价关键词检出的性能。

4.3 实验结果及分析

使用 4.1 节中的关键词检出系统作为基线系统，分别使用基于长度归一化声学置信度、关键词得分与在线垃圾得分的似然比、本文的动态排位信息得分进行关键词确认，3 种方法的 ROC 曲线如图 1 所示。

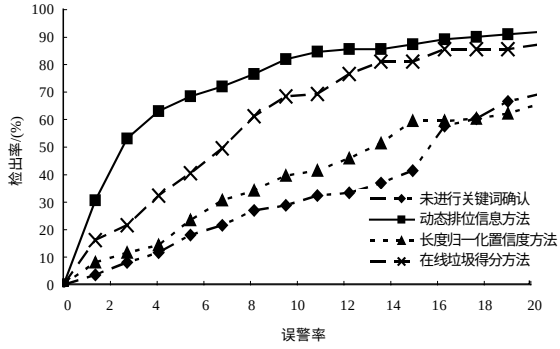

图 1　3 种方法的 ROC 曲线

可以看出，未进行关键词确认时，关键词检出系统的 FOM 为 10.9，基于动态排位信息的关键词确认方法将其提高到了 64.7，对非关键词拒识的效果明显。同时，在虚警率相同的情况下，基于动态排位信息得分关键词确认方法的检出率高于基于长度归一化声学置信度关键词确认方法以及基于关键词得分与在线垃圾得分的似然比关键词确认方法，FOM 比后两者分别高出 40.6 和 21.9。可见，本文方法的性能优于其他 2 种方法。此外，对于 44 min16 s 的测试语料，基于动态排位信息关键词确认的系统总共消耗 7 min，和基线系统基本相同。6 s 的语音不到 1 s 即可完成关键词检出，基本满足了实际应用中实时性的要求。

5 结束语

本文提出了一种适合在线垃圾模型的基于动态排位信息的关键词确认方法。它计算简便，基本不增加系统负担，易于设置统一的拒识门限，能够较好地区分关键词和非关键词，具有较好的稳定性和一定的抗噪性，是一种有效的基于在线垃圾模型的关键词确认方法。此外，本文的方法不依赖具体的关键词表，满足关键表经常变化的实际应用的需要。

(下转第 165 页)